PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Assessing Effective Mask Use by the Public in Two Countries: An Observational Study
AUTHORS	Atzema, C; Mostarac, Ivona; Button, Dana; Austin, Peter; Javidan, Arshia; Wintraub, Lauren; Li, Allen; Patel, Raumil; Lee, Dongjoo; Latham, Nathaniel; Latham, Eric; Brown, Patrick; Somogyi, Rita; Chang, Alex; Nguyen, Huong; Buerk, Sara; Chen, Bin; Zimmerman, Tristen; Funari, Trevor; Colbert, Cameron; Kea, Bory

VERSION 1 – REVIEW

REVIEWER	Lee, Albert The Chinese University of Hong Kong, School of Public Health and Primary Care
REVIEW RETURNED	24-Mar-2021

GENERAL COMMENTS	The research topic addresses important public health issue. Assessing effective mask use will have significant impact on infection control. The researchers held video meeting to ensure inter-observer reliability. However, there is no description of training of observers in data collection (assessing correct mask use and adequate physical distancing) not only ensuring inter-observer reliability and also the validity of the assessment. It is important to describe in details the training and standardisation of assessment to ensure data collection with high degree of validity. Reliability test has probably been performed but it is not reported. Readers will also be very interested to learn how the researchers validated the data collection. The readers will be interested to learn the detail methodology in selecting different sites and more description of time shift. It will be interesting to learn why Portland and Toronto are chosen for the study. There should be more information on demographic characteristics of two cities and also outbreak pattern. Why people in Toronto had lower OR of wearing mask. The study has identified interesting associating factors with mask wearing and distancing and more detail explanations would be discussed. This will enable other countries to learn from the findings in implementing mandatory mask wearing and physical distancing.

	The researchers would also consider analysis of correlation of correct mask wearing or mask wearing, compliance with physical distancing with incidence of outbreaks. The conclusion can include recommendations for public health policy.

REVIEWER	Ortelan, Naiá Oswaldo Cruz Foundation
REVIEW RETURNED	24-Jun-2021

GENERAL COMMENTS	I would like to congratulate the authors for the important research produced and very well written manuscript. Following, I make some comments: Abstract Results – page 1, lines 16-17: I suggest writing the sentence "Fourteen percent of mask-wearers wore them incorrectly" in line 13, after "Two-thirds (66.7%) were wearing a mask." Conclusions – page 2, line 1: I suggest including 'what is the consequence of this observation for the collective?' Introduction Page 4, lines 2-3: In the sentence "Public mask-use has been recommended by national and international health authorities to slow 3 the spread of COVID-19", I suggest including from which period. Page 4, line 10: I suggest writing "for an unmeasured period" than "for some time". Methods Page 7, line 22: in my opinion, the definition of mask use should be incorporated if the mask was considered adequate according to its effectiveness, otherwise, in case of using a face shield without a mask underneath, or using only a face covering such as a gaiter should be considered "incorrect type" or ate least, "incorrect". Page 8, line 11: when the subject was not alone, were both observed? This is an important component, given that 43.9% were accompanied by someone. Results Page 9, line 8: In the abstract (page 1, line 6), the authors wrote that it is a cross-sectional observational study, and in the results, the authors cited it as a cohort. However, the participants included are not monitored over time. That's why I suggest keeping "…remained in this cross-sectional study". Page 9, line 13: given that 67.7% of subjects were wearing a mask, how many used a suitable type of mask, according to its effectiveness? If the authors have this information, I think it's important to include it as additional information.

	Page 9, line 17: given the extensive confidence interval, how robust is this finding? Page 9, lines 17-18: according to the authors, “As the estimated age increased, the adjusted odds of wearing a mask did as well”, in which scenario was this observed, of mandatory mask use? In general? It would be interesting to separate these findings according to the scenario. Discussion Page 13, lines 22-23: given that the study was carried out in high-income countries, the findings cannot be applied to low- and middle-income countries either. It is important to consider that there are countries with enormous social inequality, where vulnerable/marginalized people live in environments that favor agglomerations (inside and outside their homes).

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Prof. Albert Lee, The Chinese University of Hong Kong

Comments to the Author:

The research topic addresses important public health issue. Assessing effective mask use will have significant impact on infection control.

Thank you. We agree!

1) The researchers held video meeting to ensure inter-observer reliability. However, there is no description of training of observers in data collection (assessing correct mask use and adequate physical distancing) not only ensuring inter-observer reliability and also the validity of the assessment. It is important to describe in details the training and standardisation of assessment to ensure data collection with high degree of validity. Reliability test has probably been performed but it is not reported. Readers will also be very interested to learn how the researchers validated the data collection.
As stated in the manuscript (Methods Section, Data Collection and Outcome Measures, para 3), we defined ‘correct mask use’ a priori as the following: “Incorrect mask use involved a mask with exposure of either the nares, the mouth, or both”. As per the manuscript (intro, 3rd paragraph), “adequate physical distancing” was “defined as coming within 2 metres of another person[14, 15]”. To further clarify our process, we have added the following details to the manuscript: “After several rounds of revisions, the Toronto team underwent a collective, standardized training process. The team met via recorded video conference to review ~ 30 minutes of video taken at several sites; this was conducted to minimize subjectivity that may exist in interpreting the selected metrics (e.g., correct mask usage, adherenceo physical distancing policies). Team members collectively reviewed each data point in the videos and discussed any discrepancies in interpretation or data collection as they occurred.”

We would also like to point out that observing a person with a mask on incorrectly, as per our definition (i.e. either “the nares, the mouth, or both” exposed), whether it be hanging from one ear or their nares exposed, doesn’t require extensive ‘training’, per se. Seeing the nares or mouth exposed is a simple task that persons with good vision can perform.

Rather we think of validity assessment as being important when a surrogate measure is being used to measure something else. The surrogate needs to be assessed to see whether it does indeed measure the ‘truth’ that it purports to measure. An example of assessing validity of a surrogate measure might be the assessing the validity of the triage score when a patient arrives in an emergency room: they are

assigned a score from 1 to 5 based on their sickness acuity, so the score is intended to measure the 'true' illness acuity. If the score is valid, the patient has a much more time-sensitive illness at a score of 1 than they do a score of 4. That is 'truth', but it is not so easy to confirm the validity of the triage score for this 'truth': probably higher mortality rates in the scores of 1 would be the best estimate of validity of the assigned triage score. Perhaps hospitalization, but that doesn't work for quickly reversible life-threatening illnesses (e.g. GHB overdose). But for wearing a mask incorrectly, it is a simple measure of were either nares or mouth or both exposed – it is not a surrogate for something else. Therefore we would argue that this is 'truth', no surrogate measure is needed,/span> and therefore it doesn't require a formal validity assessment. We did train all abstractors via video conference, as described above, to ensure everyone understood what they were measuring/assessing, and we also used a WhatsApp group that we used to text everyone in the group live, from data collection sites, while data collection was occurring, in order address definition issues as they arose. This ensured we had consensus on definitions. The latter has been added to the manuscript:

"After several rounds of revisions, the team met via video conference to review videos taken at several sites, to ensure inter-observer reliability in data collection by data collectors. **In addition, the study team texted each other live from the sites during data collection in order to address any uncertainties that arose around definitions via consensus; this further ensured high inter-observer reliability**." (Bold type added)

Lastly, we have added a formal assessment of interrater reliability to the manuscript: "Lastly, one member each of the Portland and Toronto teams viewed more (previously unseen) video footage taken in Toronto of 92 subjects, to determine interrater reliability using Cohen's kappa: wearing a mask 0.96, incorrect mask use 1.0.".

2) The readers will be interested to learn the detail methodology in selecting different sites and more description of time shift.

The shifts were performed during non-nighttime hours when there would be people around (i.e. there would be very few individuals on city streets in the overnight hours, so very little risk of breaches, and in stores there would be none after closure of stores at between 6pm and 9pm). Shifts were performed in 4 hour blocks during those non-nighttime hours, and data collectors were encouraged to divide their time between morning blocks (where the majority of the hours was between 08:00 and 12:00) afternoon hours (12:00-17:00), and evening hours (17:00-21:00). However given some of the data collectors were front-line workers, this was not mandatory. This has been added to the manuscript:
"Shifts were ~ 4 hours long, and were performed during non-nighttime hours (when there would be subjects present in stores, and enough elsewhere to be at risk of breaches), between 08:00 and 21:00. Each data collector was encouraged to divide their shifts evenly across those hours, and each attended at least two sites overall."

If this is not sufficient value-add for the additional word count, please let us know and we will revert to what was present in the manuscript before:
"Shifts were approximately 4 hours in duration, and each data collector went to at least two sites overall."

The different sites were selected by the group a priori, and agreed upon via consensus. The goal was to assess a variety of settings, both indoor and outdoor, with varying mandatory requirements. Currently it says the following in the manuscript:
"We chose study sites based on WHO guidelines around COVID-19 spread and mask-use (i.e. outdoors has a lower risk of spread), and anticipated differences in mask-use by site.[14,15] These included (in each city) six outdoor spaces (waterfront walkways, downtown streets, suburban business streets, public squares, parks, cemeteries), three retail stores (grocery store, drugstore/pharmacy [none in Portland], hardware store), airports (Pearson International and Portland International Airport), and public transit (bus, subway, tram)."

We have revised it to the following:

“Study sites were chosen a priori by the group via consensus, based on WHO guidelines around COVID-19 spread and mask-use (i.e. outdoors has a lower risk of spread) and anticipated differences in mask-use by site.[14,15] These included (in each city) six outdoor spaces (waterfront walkways, downtown streets, suburban business streets, public squares, parks, cemeteries), three retail stores (grocery store, drugstore/pharmacy [none in Portland], hardware store), airports (Pearson International and Portland International Airport), and public transit (bus, subway, tram).”

3) It will be interesting to learn why Portland and Toronto are chosen for the study. There should be more information on demographic characteristics of two cities and also outbreak pattern. Why people in Toronto had lower OR of wearing mask. The study has identified interesting associating factors with mask wearing and distancing and more detail explanations would be discussed. This will enable other countries to learn from the findings in implementing mandatory mask wearing and physical distancing.

The study was conceptualized by the PI (who is based on Toronto), and she sent out emails to multiple colleagues in the U.S. and Europe: individuals who publish in the PI’s main area of research, atrial fibrillation, and whom are all emergency physicians as well as scientists. This occurred in the spring of 2020, during the height of the first wave of COVID. Unfortunately, most were unable to participate due to the ongoing demands of managing COVID-19 on the front-lines, and the difficulty in quickly assembling a group of data collectors, plus the time required to train them. Fortunately Dr Kea was able to do this in a timely way in Portland, where she is based.

This is now more clearly addressed in the Limitations (in response to Reviewer #2’s comment, it has been expanded to say the following):

“Limitations of our study include the setting of two urban North American cities where the study teams were based; **our results may not apply to non-North American countries with differing governmental responses to COVID-19 and infection levels. Given enormous social inequalities both within and between countries, where vulnerable/marginalized people live in environments that favor agglomerations, our results may not apply to low- and middle-income countries**.”

We have added an appendix that describes the “demographic characteristics of two cities and also outbreak pattern”.

We have not directly addressed the comment, “Why people in Toronto had lower OR of wearing mask” because we wish to avoid overstating our results. We wonder if it was related to the infection rate (higher in Portland); however, in our study we did not specifically evaluate the myriad of factors that contribute to overall higher or lower mask use in one population over another. This would make an excellent future (perhaps qualitative) study. We have added this to the Limitations section:

“There were ~3465 new COVID cases in Toronto during our study period (population ~2,700,000) and ~ 4795 in Multnomah county, Portland (population ~650,000), raising the possibility that mask-wearing was higher in the latter city due to a higher infection rate; however, we did not formally explore reasons behind the adjusted differences in mask-wearing between the two cities. This would make an excellent future study.”

4) The researchers would also consider analysis of correlation of correct mask wearing or mask wearing, compliance with physical distancing with incidence of outbreaks.

This is an interesting suggestion. However, the study was conducted from June to August 2020, which in Ontario, was the period between the first wave and the second wave. To address the reviewer’s comment, we would need data from the first wave (March-April 2020), second wave (September 2020 – January 2021) and the third wave (April 2021 – June 2021). As the study was prospective in nature, this is not feasible for us to provide.

The conclusion can include recommendations for public health policy.
Thank you for this suggestion. We have added the following to the Conclusion (bold type), in addition to the final statement that is a recommendation for public health policy:

"Elderly persons were the most likely to make mask-wearing errors**, and therefore should be targeted by educational mask-wearing campaigns**. A mandatory requirement to wear a mask was the greatest predictor of both mask-wearing and correct wear, and was not associated with an inadvertent increase in breaches. These results support mandating mask-use in public settings as an effective **public health** strategy to prevent the spread of COVID-19."

Reviewer: 2

Dr. Naiá Ortelan, Oswaldo Cruz Foundation
Comments to the Author:
I would like to congratulate the authors for the important research produced and very well written manuscript.

Thank you.

Following, I make some comments:

Abstract
Results – page 1, lines 16-17: I suggest writing the sentence "Fourteen percent of mask-wearers wore them incorrectly" in line 13, after "Two-thirds (66.7%) were wearing a mask."

Thank you for this suggestion to improve the flow: this has been done.

Conclusions – page 2, line 1: I suggest including 'what is the consequence of this observation for the collective?'

The line 1 on page 2 states, "Mandatory mask use was associated with a large increase in mask-wearing." Increased mask use is the consequence (more specifically, the association, since without randomization one cannot prove causation). Given that this is an observational study, we are reluctant to delineate the consequences (presumably less COVID transmission and cases), since we cannot prove causality using our study design. We feel that is important not to overstate our study conclusions.

Introduction
Page 4, lines 2-3: In the sentence "Public mask-use has been recommended by national and international health authorities to slow 3 the spread of COVID-19", I suggest including from which period.

Thank you – this has been added: "Public mask-use was recommended in spring 2020 by national and international health authorities in order to slow the spread of COVID-19". The dates in Canada were May 2020 and by the CDC for America in April 2020 (i.e. well before our study period).

Page 4, line 10: I suggest writing "for an unmeasured period" than "for some time".

This has been changed, as per the reviewer's excellent suggestion.

Methods
Page 7, line 22: in my opinion, the definition of mask use should be incorporated if the mask was considered adequate according to its effectiveness, otherwise, in case of using a face shield without a

mask underneath, or using only a face covering such as a gaiter should be considered “incorrect type” or ate least, “incorrect”.

Thank you for highlighting this deficit in our definitions. The use of a face shield without a mask underneath was **not** considered a mask at all in our study, as the nares and mouth were exposed. We have clarified this in the Methods section.

As for cloth masks, the authors of this study evaluated the suitability of a mask in accordance with public health guidelines from both countries (CDC and Public Health Agency of Canada, for America and Canada, respectively) **at the time of the study**. At that time, these permitted a wide range of masks or face coverings including medical masks, non-medical cloth masks, and scarves or gaitors that covered the nose and mouth. I.e. these were “considered adequate”. We have added the following sentence to this section: “For the purpose of our study, **consistent with guidelines issued at the time of the study from both countries,**[1,2] a mask was defined as either a surgical mask, N95 respirator, cloth mask, a gaiter, and a cover over a baby stroller. **A face shield worn without a mask was considered ‘no mask’**.” (Bold type added to addition)

Since the public health agencies considered these masks “adequate” **at the time of the study period**, we did not collect specific data on the use of cloth or gaiter masks. We argue that the subjects were doing what was considered appropriate at the time, and would therefore be very likely to use the correct/recommended masks in future, as guidelines change. If the study had also collected infections, use of masks that were later deemed less effective in the pandemic would be necessary to correlate with infection rate, but we did not assess these; rather we were only interested in how well subjects in each city a) wore masks (that were officially approved by public health), and how often they wore them incorrectly (according to public heath instructions)

Page 8, line 11: when the subject was not alone, were both observed? This is an important component, given that 43.9% were accompanied by someone.

Yes, when subjects were together in a group they were all assessed.

Results

Page 9, line 8: In the abstract (page 1, line 6), the authors wrote that it is a cross-sectional observational study, and in the results, the authors cited it as a cohort. However, the participants included are not monitored over time. That’s why I suggest keeping “…remained in this cross-sectional study”.

The opening statement, “After removal of 26 (0.07%) subjects who did not have their mask-use recorded, 36,808 individual observations remained in the cohort” has been changed to “After removal of 26 (0.07%) subjects who did not have their mask-use recorded, 36,808 individual observations remained in this cross-sectional study.”

Page 9, line 13: given that 67.7% of subjects were wearing a mask, how many used a suitable type of mask, according to its effectiveness? If the authors have this information, I think it's important to include it as additional information.

Please see the answer above, about public health agency approved masks, under “Page 7, line 22”.

Page 9, line 17: given the extensive confidence interval, how robust is this finding?

The wide CI is primarily a function of the large effect (OR 79.2, corresponding to a log-OR of 4.4 and a moderate standard error). The CI for the log-OR is narrower (47.4-135.1, corresponding to a log-OR of 3.9-4.9). When this is exponentiated, it amplifies the width of the CI by the nature of the exponential function. The finding is indeed robust.

Page 9, lines 17-18: according to the authors, “As the estimated age increased, the adjusted odds of wearing a mask did as well”, in which scenario was this observed, of mandatory mask use? In general? It would be interesting to separate these findings according to the scenario.

We agree that the reviewer’s suggestion is interesting. However, we did not have an a priori hypothesis to assess this, and a post-hoc assessment would be less than rigorous. As well, due to the low number of subjects in the 80+ category within each of the different locations (e.g., 29 in outdoor spaces: please see Table 1), we are concerned about the potential instability of the fitted models and that associated CIs would be very wide. (The low number of 80+ year-olds was not surprising as they were a high risk group and advised to stay indoors/not venture into public spaces where they might contract COVID-19). For these reasons, we have not added these analyses.

Discussion

Page 13, lines 22-23: given that the study was carried out in high-income countries, the findings cannot be applied to low- and middle-income countries either. It is important to consider that there are countries with enormous social inequality, where vulnerable/marginalized people live in environments that favor agglomerations (inside and outside their homes).

Thank you for this insightful suggestion. We have added it here, adjusting the sentence as per Reviewer #1’s suggestion as well: “Limitations of our study include the setting of two North American cities; **our results may not apply to non-North American countries with differing governmental responses to COVID-19 and infection levels. Given enormous social inequalities both within and between countries, where vulnerable/marginalized people live in environments that favor agglomerations, our results may not apply to low- and middle-income countries**.”(bold type added to addition)

Reviewer: 1

Competing interests of Reviewer: I have no competing interests.

Reviewer: 2

Competing interests of Reviewer: None.

VERSION 2 – REVIEW

REVIEWER	Ortelan, Naiá Oswaldo Cruz Foundation
REVIEW RETURNED	20-Sep-2021

GENERAL COMMENTS	I would like to once again congratulate the authors for the work they have done, and thank them for incorporating my suggestions into the article. I would like to point out again that perhaps the authors can highlight the limitations that given the extensive confidence

	interval of the main result (mandatory mask use overwhelmingly associated with wearing a mask: OR 79.2; 95% CI 47.4-135.1), how robust is this finding? Moreover, I hope it will be a successful publication that inspires other works, given the worldwide context of COVID-19.